ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE1BC002

# DATA MANAGEMENT INFRASTRUCTURE FOR EUROPEAN XFEL

J. Malka*, S. Aplin, D. Boukhelef, K. Filippakopoulos, L. Maia,
T. Piszczek, G. Previtali, J. Szuba, K. Wrona
European XFEL, Schenefeld, Germany
S. Dietrich, M. Gasthuber, J. Hannappel, M. Karimi, Y. Kemp, R. Lueken,
T. Mkrtchyan, K. Ohrenberg, F. Schluenzen, P. Suchowski, Ch. Voss
Deutsches Elektronen-Synchrotron (DESY), Hamburg, Germany

*Abstract*

Effective data management is crucial to ensure research data is easily accessible and usable. We will present the design and implementation of the European XFEL data management infrastructure supporting high-level data management services. The system architecture comprises four layers of storage systems, each designed to address specific challenges. The first layer, referred to as Online, is designed as a fast cache to accommodate extremely high rates where up to 15 GB/s of data is generated during experiments with a single scientific instrument. The second layer, called high-performance storage, provides the necessary capabilities for data processing both during and after experiments. The layers are incorporated into a single InfiniBand fabric and connected through a 4.4 km long 1 Tb/s link. This allows fast data transfer from the European XFEL experiment hall to the DESY computing centre. The third layer, mass storage, extends the capacity of the data storage system to allow mid-term data access for detailed analysis. Finally, the tape archive provides data safety and a long-term archive (10+ years). The high-performance and mass storage systems are connected to computing clusters. This allows users to perform near-online and offline data analysis, or alternatively export data outside the European XFEL facility. The data management infrastructure at the European XFEL has the capacity to accept and process up to 2 PB of data per day, which demonstrates the remarkable capabilities of all the sub-services involved in this process.

## EUROPEAN XFEL FACILITY

The European XFEL Facility is one of the most advanced sources of pulsed, extremely intense and coherent radiation in the hard and soft X-ray range. These properties of X-ray flashes attract various research communities and scientists, who use them in a diversity of scientific investigations. The facility spanning a length of 3.4 km, extends from the DESY campus in Hamburg to the town of Schenefeld situated in Schleswig-Holstein, where the European XFEL experiment stations are located. Currently, a scientific experiment can be performed using one of the seven available instruments and utilising one of the three self-amplified spontaneous emission light sources called SASE. The layout of scientific instruments and SASEs is illustrated in Fig. 1.

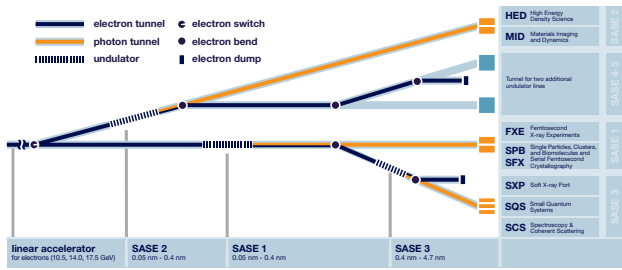


Figure 1: Beamlines and instruments of the European XFEL

### *Scientific Data Policy*

The European XFEL Scientific Data Policy defines the rights and responsibilities for data of all parties involved in experiments and sets rules for data management. It was approved and published in 2017, shortly before the first user experiments were performed. Based on the policy, a set of data services has been created together with underlying storage and computing infrastructure. Through the implemented services, scientists are able to catalogue and later find the data, assess its quality and trigger some actions like, for example, data archiving, and specialised data processing pipelines. The data policy is currently under revision in order to be aligned with the FAIR data principles. It will also address the challenges related to the vast data volumes by introducing data management plans, data reduction concepts and a new data retention policy. Along with the updated policy, the data services are being reviewed with the goal of implementing new data handling concepts.

### *Meta Data Catalogue*

Data generated in the context of scientific experiments at the European XFEL facility is curated with the help of metadata catalogue service (myMdC) [1]. The metadata catalogue service besides storing information about experiments data, acts as a hub for integration with other services (e.g. data acquisition system, electronic logbook, calibration service, storage infrastructure) and is able to execute various data management workflows. The myMdC is built on top of the relational database, which links experiment meta-data and data together. It allows scientists to define experiment techniques, measurements and sample types for each experiment dataset. Through extensive use of microservices, it allows the automation of data infrastructure tasks such as creating an underlying file system structure for each experiment and defining data access roles within the experiment team. It is also used to trigger data migration processes, ini-

* janusz.malka@xfel.eu

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE1BCO02

tialize data processing tasks and track locations of data. The myMdC could be accessed as a web application or through the RESTful APIs.

### *Data Volume*

Since the beginning of the facility operation, an enormous increase in the generated raw data volume has been observed (see Fig. 2) reaching, as of the time of publication the size of 102 PB. In addition, about 50 PB of storage is used to keep the processed data. The following sections will describe the data infrastructure which is able to accept and process this huge amount of data, and deal with data rates as large as a record of 7 PB generated in a single week (end of November 2021).

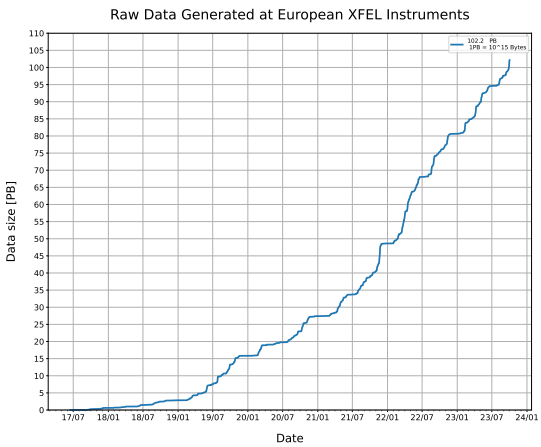


Figure 2: Size of raw data volume generated by European XFEL instruments in function of time.

## DATA ARCHITECTURE MODEL

The data architecture model applied in European XFEL is similar to a cascade waterfall with data lakes on each level. In the first layer, called online, the computing and storage resources that are located near experiment stations are used exclusively to support a single experiment during the assigned time. The second layer, called high-performance storage, is common for all experiments and provides resources for fast data processing during the experiments (a couple of days) and shortly after it. The third layer, mass storage, extends the capacity of the high-performance storage and is typically used to store data of past experiments until the data is being analysed. The last layer is the tape archive, which provides resources for long-term data preservation. All mentioned layers, are shown in Fig. 3, and will be discussed in more detail in the sections below.

### *Online*

Each SASE is equipped with dedicated data infrastructure. Since, at a given point in time, only one experiment is scheduled per SASE, exclusive access to the computing and storage resources is provided for all active experiments. The

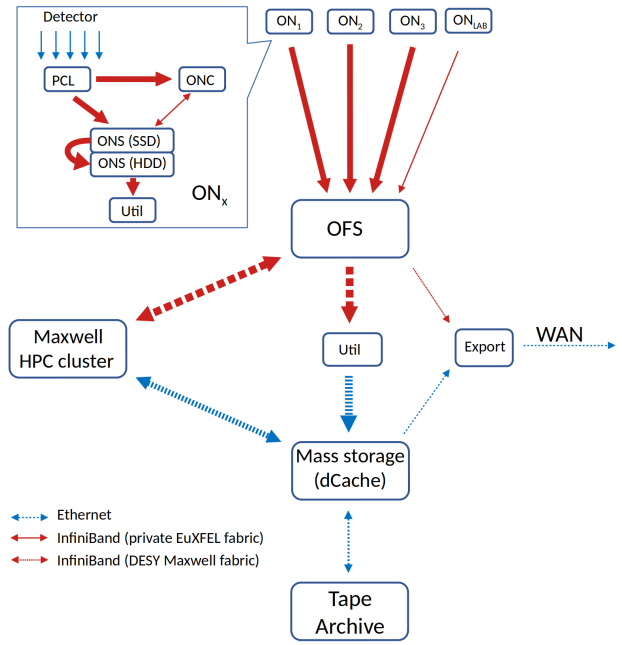


Figure 3: Storage and computing infrastructure diagram. The red arrows show the direction of the data flow in EuXFEL private InfiniBand fabric, the doted red arrows represent the data flows in DESY Maxwell InfiniBand fabric and blue arrows data traffic over Ethernet. PCL - PC-Layer cluster, ONC - Online Computing cluster, ONS - Online Storage cluster, OFS - High-performance Offline Storage, Mass storage (dCache), Tape archive, and Maxwell HPC cluster are indicated by the blue rounded rectangles, together with utility nodes performing administration tasks and actual data movement.

core of the online environment is the Online Storage (ONS) cluster. Each ONS cluster is based on IBM Storage Scale System [2], which consists of two sub-systems: one built on top of fast SSD drives and the second utilising HDD drives. The former is capable of high performance, which is suitable for accommodating bursts of IO operations, whereas the latter has a capacity to store data for at least a full experiment day. There are a few file systems on the ONS cluster, and each of them is configured for different use cases as home folder, software repository, calibration constants storage and finally the space for experiment raw data and intermediate results of online or near-online data processing and analysis. The last mentioned file system holds three storage pools. The first one, the system pool, hosts the meta-data of the file system and uses SSD drives as the underlying hardware. The second is the cache pool, which hosts the most recent experiment data and is also built on top of SSD drives, thus allowing very fast data ingest and read access. The third is the data pool, that hosts colder experiment data and is built out of HDD drives. The automatic migration process effectively drains data from the cache to the data pool before the cache pool is filled up. Apart from that, the raw data collected during the experiment is copied to the high-performance storage cluster at the DESY computing centre. At a single

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE1BCO02



point in time, those three processes: data ingest, cache-data pool drain and data copy to high-performance storage are competing for free IO resources. The ONS clusters prioritize the data ingest above the other two processes by limiting the available IOPS. The typical data rates are: 15 GB/s of data ingest, 20 GB/s of data drain, and 30 GB/s of data copy, measured simultaneously. The file systems hosted on the ONS cluster are exported to two client clusters: the so-called PC-Layer (PCL) cluster and the Online Computing (ONC) cluster. The PCL cluster allows Data Acquisition (DAQ) system to aggregate and store experiment data on the ONS cluster. It consists of many physical servers, each running one or more Data Aggregator (DA) processes. DAs create HDF5 [3] files which contain data and metadata generated by large 2D detectors and other devices and sensors. In addition to the data writing, a DA can stream data via InfiniBand network to the ONC cluster. The ONC cluster consists of several GPU and CPU-only nodes. If required, preliminary data corrections are performed on the ONC cluster nodes and that allows live data preview and fast feedback analysis.

### *High-Performance Storage (Offline)*

Much larger than the online storage cluster, the high-performance storage, called also the offline storage cluster (OFS), is used to store the raw data as well as processed data for detailed analysis and plays an intermediate role in data transfer from the Online storage to the mass storage system. This storage cluster is build on top of IBM Storage Scale System building blocks, small fast ones equipped with NVMe drives for good metadata performance and large HDD based ones for capacity. Due to the number of building blocks (currently 11 capacity blocks and two NVMe blocks) a high degree of parallel data access allows for very good performance, a read performance of more than 175 GiB/s was demonstrated.

This cluster is evolving over time in the sense that old building blocks, that have outlived the support contracts are removed from the cluster and new building blocks are added to compensate the removed capacity and also to expand the available space as demand grows over time, thus creating a cluster that combines several generations of the storage hardware[1].

Once data arrive from the ONS cluster, an archival process is triggered, that copies data from the OFS storage to the mass storage, whence data will automatically be copied once again to the tape library.

A check process then records for each file the number of disk and tape copies, so that the valuable space in the OFS cluster can be freed, while being assured that two copies of the data exist.

### *Mass Storage*

The mass storage infrastructure is the third tier in the data-management architecture of the EuXFEL facility and is based on the dCache technology [4]. It is currently the largest by capacity as well. Its main role is to provide long-term storage for raw data and part of processed data as well as input storage to archive these data on tape. In contrast to the Online and Offline clusters, it is based on commodity hardware using Ethernet network as well as Nearline SAS disks.

At its core, dCache is based on a micro-service architecture with components managing the storage servers, the client access as well as handling internal services for authentication and authorization, name-space, interaction with the tertiary storage systems, as well as storage server selection. By itself dCache does not handle the underlying storage, rather it relies on the underlying infrastructure to provide resilience against hardware failures. However, it can be configured to store multiple copies of files on different storage nodes. Data-access is granted via different authorization schemes, e.g. X.509 certificates or OpenID connect providers, as well as through multiple protocols, such as FTP, XrootD, WebDAV, NFS4.1 and dCap, native to dCache itself. At the EuXFEL currently only the latter two see major usage. As dCache exposes a unified name-space, the view on the file system is uniform across all protocols.

Each storage server, referred to as a pool, is independent of all others, allowing easily to scale out horizontally. The overall capacity can be increased by simply adding new pools. In Fig. 4 the development of the dCache capacity is shown. For reference, the development of the capacity for the Particle Physics experiments at DESY is added.

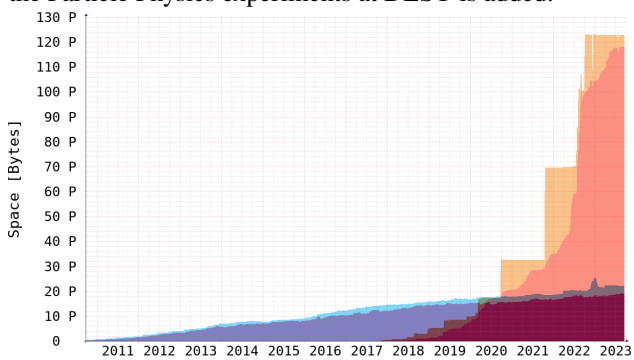


Figure 4: Development of the EuXFEL dCache storage capacity (orange) at the DESY Computing Centre compared to capacity used by the Particle Physics communities (blue) over the last decade. The darker shades indicate used space; lighter shades the overall space.

The instance for EuXFEL is split logically into two major parts, a disk-only and one that is connected to the DESY tape infrastructure. The first is designed for users to store additional files needed in their analyses. Each file is replicated to another storage server in order to prevent data loss in case of a total failure of one storage server. Overall, the capacity of the disk-only area is about 1%. The second part is used to store the raw data files. Each of these pools is connected to the DESY tape infrastructure following a classical HSM setup. The main difference to a classical HSM setup is,

[1] Currently 3 building blocks with Power8 CPUs, 6 ESS5000 building blocks and 2 ESS3500 units, along with 2 ESS3000 models for the meta data

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE1BCO02

that users cannot trigger restores themselves. Clients using e.g. NFS-4.1 will receive a permission denied error when accessing on tape-only file. Stages are triggered either by the EuXFEL data managers or the DESY dCache operations team, after previews assessment

Based on specific tags assigned to the directories in which the file are saved, each file is assigned a specific storage class, e.g. the detector and run period, and is migrated to tape. These storage classes allow the underlying tape system to store the data logically connected. This means files from one user experiment are grouped logically on a physical tape set for efficient usage of space and later restores.

File locality in general is managed by using Quality of Service (QoS) classes as implemented in dCache. By default, all raw data files are set to be on disk and tape, hence the QoS level is set to disk+tape. In case space needs to be reclaimed, the EuXFEL data managers can modify the QoS level to tape, which will make the files cached and allow dCache to remove them from disk if needed. In case these files are requested again by scientists, changing QoS back to disk+tape will trigger a restore and will ensure the files remain on disk. These transitions are triggered through the dCache REST-API.

### *Tape Archive*

EuXFEL uses tape storage for long term archival of data. Tape comes with several benefits such as capacity, price, durability and energy efficiency. The IBM TS4500 tape library, provides storage through 12 TB LTO8, 18 TB LTO9 and 20 TB 3592JE cartridges with transfer rates of about 400 MBps per drive. Around 105PB of EuXFEL data has been archived to tape. The tape archive backend is provided by the CERN Tape Archive (CTA) [5] software, which has seamlessly been integrated with dCache.

### *Maxwell High-Performance Computing Cluster*

Data analysis, processing and simulations related to experiments at the EuXFEL are almost exclusively performed on the Maxwell HPC cluster. The cluster has all the ingredients of a typical HPC platform with SLURM-scheduling, low-latency InfiniBand backbone, cluster file-systems and a total of 940 compute nodes with a theoretical peak performance of 4000 TFlops.

The Maxwell cluster is however - in contrast to conventional HPC systems - extremely heterogeneous, thereby reflecting the heterogeneity of the DESY campus, which hosts a very large number of independent institutions covering a wide area of scientific fields. Rather than letting each institution operate their own compute island, the Maxwell cluster allows bringing in institutional compute resources, which are conveniently shared by all users of the Maxwell cluster. This cooperative model leads to a good resource utilization allowing users to allocate a substantially larger number of compute nodes than an institutional platform could offer, but comes at the price of a heterogeneous compute environment.

EuXFEL is one and by far the largest member of the Maxwell platform, contributing 439 CPU-only and 30 GPU nodes or roughly 1200 TFlops which account for about 30% of the cluster's compute power. The compute nodes are essentially distributed over two general partitions serving staff members and users, and three partitions dedicated to compute tasks and calibrations during user experiments. The partitions are overlapping and can be reserved dynamically to cope with the needs of running experiments.

The primary focus of the Maxwell cluster to serve primarily photon science data processing leads to a somewhat unusual HPC workload distribution. Out of roughly 4 Million batch jobs per year, more than 90% of the jobs are actually single-node jobs. Jobs running in the EuXFEL context contribute for more than 50% of jobs and quite remarkably for almost all many-node jobs (jobs with more than 16 nodes). Serving running experiments with the aim to enable large scale data processing requires immediate availability of sufficient compute resources. Guaranteed immediate access to compute resources is not perfectly compliant with the scheduling configuration (FIFO+Backfilling). Compute resource for running experiments are hence allocated from highly prioritized partitions, using node-reservations to distribute resources across concurrent experiments according to individual computational requirements.

The Maxwell cluster also offers a customized JupyterHub instance, which allows selecting SLURM partitions and reservations to launch Jupyter server as regular batch jobs in customizable environments. Not surprisingly, Jupyter notebooks became hence quite popular among scientific users: more than 500 of the 2800 Maxwell users occasionally launch Jupyter notebooks. However, the number of Jupyter jobs account for only 1% of all batch jobs, and the consumption of compute resources is «1% in average, which is in view of the typically very poor resource utilization by Jupyter notebooks quite fortunate. However, an increasing number of use cases utilize Jupyter notebooks for online analysis during experiments, which makes the Jupyter ecosystem a critical infrastructure ingredient.

### *Sustainability*

In addition to the ongoing research into increasing computing efficiency, which benefits from complementary activities in different scientific communities, notably Particle Physics and HPC, recent activities have been focusing on raising awareness by generating summaries, per user and job, about power consumption and related greenhouse gas emissions. This is currently in preparation for the Maxwell HPC cluster. By having a single shared compute cluster for the non matching tasks (data processing during experiment vs. classical batch) described in detail in the previous section, we avoid installations of large scale dedicated resources leading to lower efficiency in utilization and power. In addition, a more complete and detailed power metering for all data centre equipment is planned. Dedicated activities will start in 2024 within a project supported by external funding.

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE1BCO02



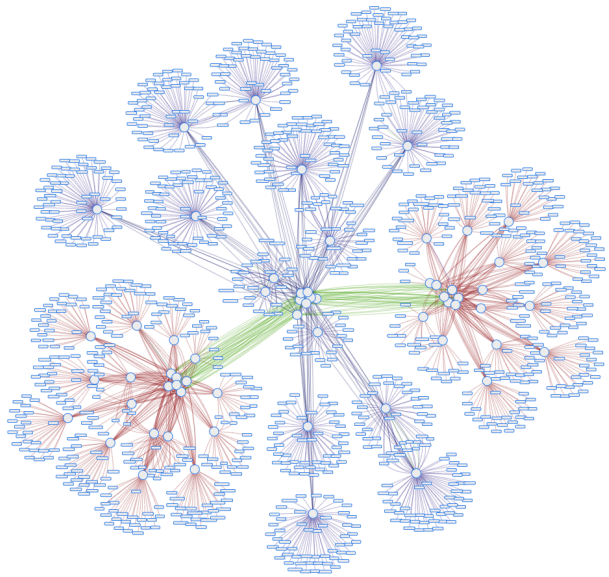

Figure 5: InfiniBand fabric of the Maxwell cluster. An HDR (violet) root layer connects two EDR (green) layers (blobs at top and bottom), that then connect nodes via FDR (red) leaf switches. The smaller "branches" are HDR switches that connect nodes via HDR100 splitter cables

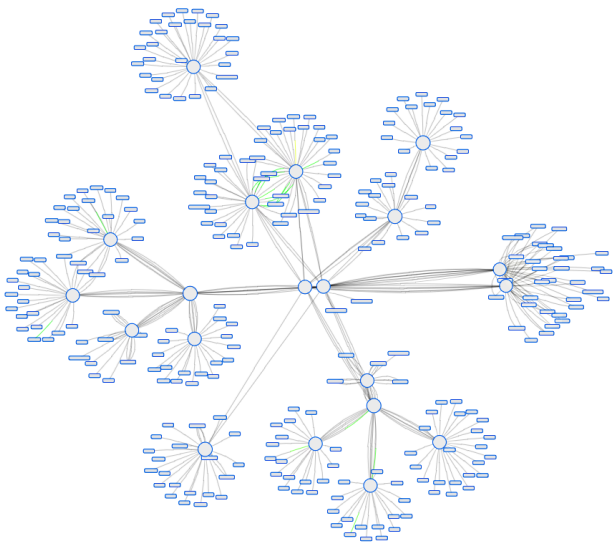

Figure 6: Private online InfiniBand fabric. The blob on the right side, connected with 20 lines to the centre is locate at the DESY Computing centre and serves the OFS cluster. The other big branches represent the hardware for each SASE, with the two-stage part at 2 o'clock being the detector lab and the small branch at 7 o'clock the storage cluster for that. The lines between the centre and the OFS cluster blob represent the long-range connection between the EuXFEL and DESY premises. Each line is a 4.4 km EDR link, where the effective throughput is about 50 Gb/s, due to the buffering done in the switch on the sender side, giving 1 Tb/s in total as required.

### *Network*

The IBM Storage Scale clusters, i.e. the ONS and OFS storage clusters as well as the Maxwell compute cluster and the online compute and PCLayer clusters all use InfiniBand as fast and low latency networking fabric with RDMA[2] enabled. To insulate the time-critical online world dedicated to data acquisition from the shared Maxwell cluster with several scientific communities, two independent InfiniBand fabrics are used: DESY Maxwell (Fig. 5) connects the Maxwell nodes and all the OFS storage nodes, one private fabric (Fig. 6) connects the online clusters and also the OFS storage nodes, so that the OFS storage nodes form the connection between the online and offline worlds.

Similarly as the storage cluster, also the InfiniBand is a 'rolling' installation. As new components are added over time, it contains several generations of InfiniBand, ranging from FDR over EDR to HDR. With the advent of adaptive routing that uses the least loaded link between switches especially the private fabric could be used much more efficiently. There are no more congested links slowing down network traffic, as was the case before this feature was available.

Although the main traffic over the InfiniBand fabrics is file I/O, there are several other applications. On the private fabric data is transferred directly from the PCLayer nodes to the online compute nodes to provide a fast and low-latency data preview and in the Maxwell fabric MPI jobs also use the InfiniBand network for the inter-node communication. Administration tasks, user login and dCache communication are performed via Ethernet.

### *Monitoring*

In order to detect and minimize outages and provide guidance for future scaling requirements, service and performance monitoring is applied to the infrastructure.

A central Icinga 2 [6] installation in DESY computing centre is responsible for service monitoring of the EuXFEL data management infrastructure. The Icinga 2 Agent executes service checks to determine the health of the service and reports the result to the Icinga 2 instance. The following service checks monitor the infrastructure:

- System Health: CPU, memory and disk usage checks
- Hardware Health: Memory, power supply and hard drive failure checks
- IBM Storage Scale Health: Daemon availability, long running waiters (Deadlock detection), pool usage

Service checks are written in various programming languages (C, Bash, Perl, Python).

Any hardware related failures during the normal business hours are handled by the local operation team. Critical services failures outside business hours are handled by on-call duty personnel. More complex failures, e.g. Storage Scale related interruptions, are escalated to the storage administration team.

For performance metrics gathering, two different tools are used:

- IBM Storage Scale Performance Monitoring Tool
- Telegraf for metric collection and Graphite for storing numeric time-series data

[2] Remote Direct Memory Access

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-WE1BCO02

The IBM Storage Scale Performance Monitoring Tool consists of 2 components:

- Collector: Store and query of IBM Storage Scale metrics
- Sensor: Collect IBM Storage Scale metrics from a host

Sensors are collecting Storage Scale related metrics, e.g. I/O throughput and latency, IOPS, block usage for filesets and pools and waiters. These metrics are stored by the collector and available for queries.

Telegraf [7] is responsible for collecting several standard Linux metrics, like CPU, memory and disk usage metrics, throughput of network interfaces and also the power consumption. While there is an overlap between the Icinga 2 system health checks and system metrics, they are stored in different databases with different capabilities. The Graphite [8] time-series database is specialized for numeric data and supports aggregation of metric values. This allows to reduce the storage consumption at the price of reduced precision for historic values.

In order to visualize all collected metrics, Grafana [9] is used. Although Graphite and IBM Storage Scale Collector have their own visualization capabilities, using Grafana is advantageous. Grafana allows visualizing metrics from both time-series databases on a single dashboard. While Graphite is natively supported by Grafana, IBM Spectrum Scale Bridge for Grafana [10] is required to visualize IBM Storage Scale metrics. With the bridge, OpenTSDB queries from Grafana are translated to the query language of the collector.

## CONCLUSION

The infrastructure and services presented in this paper have been in production since 2017. Since then, continuous optimisation and automation have reduced the administrative burden of running the system. In parallel, the proven rate and capacity at which detector data could be stored and processed were increased by a factor of 2-3, closely following the scaling of the technology used. Current technology roadmaps and industry co-operations indicate that future scaling is achievable without architectural changes. One of the key components has been the long-range (4.4 km) InfiniBand connection between ONS and OFS pushing detector raw data through many physical links using RDMA capable protocols and utilizing standard HDR InfiniBand switches.

## ACKNOWLEDGEMENTS

We wish to acknowledge the help provided by the instrument scientists and data experts of European XFEL GmbH and DESY-IT colleagues not mentioned in the author lists. We would also like to show our deep appreciation to our business partners who are helping us provide an excellent data service for users of our facility.

## REFERENCES

[1] myMdC, https://in.xfel.eu/metadata/

[2] Frank Schmuck and Roger Haskin, "GPFS: A Shared-Disk File System for Large Computing Clusters", in *Proc. FAST 2002 Conf. File Storage Technol.*, Monterey, CA, USA, 2002.

[3] The HDF Group. Hierarchical Data Format, https://www.hdfgroup.org/HDF5/

[4] Tigran Mkrtchyan *et al.*, "dCache: Inter-disciplinary storage system", in *25th Int. Conf. Comput. High Energy Nucl. Phys. (CHEP 2021)*, vol. 251, 2021, p. 02010. doi:10.1051/epjconf/202125102010

[5] The CERN Tape Archive, https://cta.web.cern.ch/cta/

[6] Icinga, https://icinga.com/

[7] Telegraf, https://www.influxdata.com/time-series-platform/telegraf/

[8] Graphite, https://graphiteapp.org/

[9] Grafana, https://grafana.com/

[10] IBM Spectrum Scale Bridge for Grafana, https://github.com/IBM/ibm-spectrum-scale-bridge-for-grafana